8-26-2004

Distributing and Synchronizing Heterogeneous Metadata in Geospatial Information Repositories for Access

Elaine L. Westbrooks
University of Nebraska-Lincoln, elainelw@email.unc.edu

Published in METADATA IN PRACTICE, edited by Diane I. Hillman & Elaine L. Westbrooks (Chicago: American Library Association, 2004).

9 Distributing and Synchronizing Heterogeneous Metadata in Geospatial Information Repositories for Access

Elaine L. Westbrooks

IN 1998 THE Albert R. Mann Library created the Cornell University Geospatial Information Repository (CUGIR), a web-based repository providing free access to geospatial data and metadata for New York State.[1] Since its inception, CUGIR has undergone a series of changes and upgrades in response to emerging standards and technologies in the field of geospatial information systems (GIS) and digital library research. Its continuous adoption of new library and GIS standards and developments has made CUGIR increasingly more accessible to users within Cornell University and beyond.

The Cornell University Geospatial Information Repository has a number of characteristics that pose unique challenges for digital library developers. First, most GIS repositories manually distribute data and metadata via CD-ROM, whereas CUGIR freely distributes data and metadata via the World Wide Web, making it a true digital library. Second, it is rare to have a geospatial repository whose invention, support, and subsequent development occur within an academic research library. Academic GIS repositories or units are typically under the jurisdiction of urban planning, architecture, or geography departments. Because CUGIR is positioned in a library environment, it embraces standards and practices associated with the preservation, retrieval, acquisition, and organization of information. The library community has always been concerned with the archiving and

version control of information, and believes that consistent application of standards will increase interoperability. The library community also believes that metadata, though costly and difficult to create and manage, adds value to whatever it describes. The GIS community is most concerned with creating data efficiently, easing the burden of metadata, and distributing data according to user requests. Generally speaking, GIS data are qualitatively different and more problematic than most digital library objects, including moving images.[2] More importantly, perpetual updating, versioning, and "editioning" of data at the owner's request makes GIS data management and metadata management difficult.[3] CUGIR reserves a position in two communities, library and GIS, requiring the CUGIR team to embrace the standards of both.

This sum of CUGIR's unique characteristics led the team to ask the following questions: if one were to create a perfect and heterogeneous metadata management system for a digital library, like CUGIR, what characteristics would it possess? How would it behave? What problems would it solve? The CUGIR team set out to create a system characterized by automatic metadata updating and digital object permanence. The system would be designed to behave in a predictable fashion, reduce work and costs, and increase access. The CUGIR metadata model is not a perfect metadata management system, but it is efficient. This is largely because it is a hybrid system embracing the standards, research, and practices of the library community while adopting the GIS community's most attractive feature, its software.

In striving for metadata management perfection, the CUGIR team became keenly aware of the shortcomings in the way GIS software manages digital objects and metadata, primarily the lack of version control for objects and preservation for metadata. Subsequently, these shortcomings were examined under the lens of the Functional Requirements for Bibliographic Records (FRBR) conceptual data.[4] This set of requirements was sponsored by the International Federation of Library Associations and Institutions' (IFLA's) section on cataloging to address the changes in cataloging processes. The FRBR addresses three groups of entities, but for CUGIR's purposes the first group, which outlines the primary relationships between works, expressions, manifestations, and items, is most critical. In particular, FRBR's use of the concept *work* was examined in the context of CUGIR, and it was through this lens that the team began to view the differences among metadata surrogates or *entities* within CUGIR.

Similarly, the weaknesses of the typical digital library metadata model, particularly its disregard for automation, were addressed in two ways. First, the storage of surrogate records for multiple *manifestations* of the same *expression* was eliminated. Second, the automatic metadata-creation tools unique to GIS software applications were exploited to increase efficiency. These changes proved to be a step in the right direction toward improved management of heterogeneous metadata.

The purpose of this chapter is to introduce the CUGIR metadata management model, whose primary goal is access. This model specifically attempts to address the following problems that can hinder access:

1. Management of multiple metadata schemas, i.e., FGDC, MARC, and DC, that occur in multiple manifestations and expressions in CUGIR
2. The lack or absence of fixity and persistence or permanence of geospatial digital objects[5]
3. The creation and maintenance of metadata that is typically difficult, costly, and time-consuming
4. The lack of tools to automate the creation and management of metadata, in particular, metadata synchronization

It was the goal of the CUGIR team to take the best of both worlds (digital libraries and GIS applications) and merge them to make a powerful system from which both communities could benefit. Although this model was chiefly designed for geospatial data and metadata, it can be applied to other types of digital libraries.

BACKGROUND

CUGIR is a clearinghouse and repository that provides unrestricted access to geospatial data and metadata, with special emphasis on those natural features relevant to agriculture, ecology, natural resources, and human-environment interactions in New York State. Staff at the Albert R. Mann Library of Cornell University began looking at ways to disseminate geospatial data from Mann's collections via the Web in 1995, and in 1998 they established a web-based clearinghouse for New York State geospatial data and metadata. Building a clearinghouse entailed creating partnerships with local, state, and federal agencies; understanding how to interpret and apply the Federal Geographic Data Committee (FGDC) Content Standard

for Digital Geospatial Metadata (CSDGM); and designing a search and retrieval interface, as well as a flexible and scalable data storage system.[6]

The CUGIR team consists of five regular members, each coordinating work within their areas of specialty. Primary responsibility for the overall coordination of clearinghouse development rests with the GIS librarian. This team provides for the management, preservation, organization, and storage needs of datasets that are distributed in CUGIR, but which are owned by various departments in New York State governmental agencies as well as Cornell-affiliated departments, agencies, and researchers.[7] Although the CUGIR team strives to make access better, the biggest responsibility of the team is adding value to the data within CUGIR.

The Cornell University Geospatial Information Repository is one of 250 international nodes within the National Geospatial Data Clearinghouse that contain searchable metadata records describing geospatial datasets. All nodes are located on data servers using the Z39.50 information retrieval protocol. As a result, nodes can be linked to a single search interface where the metadata contents of all nodes, or any subset in combination, can be searched simultaneously. The Cornell repository, like most clearinghouse nodes, has its own website with customized browsing and searching interfaces. Usage statistics indicate that CUGIR's utility and popularity continues to grow. Since 1998, CUGIR data requests have increased by at least 40 percent each year. In fact, it is projected that CUGIR will record over 100,000 requests in 2004, the most for any single year since the repository was established in 1998.[8]

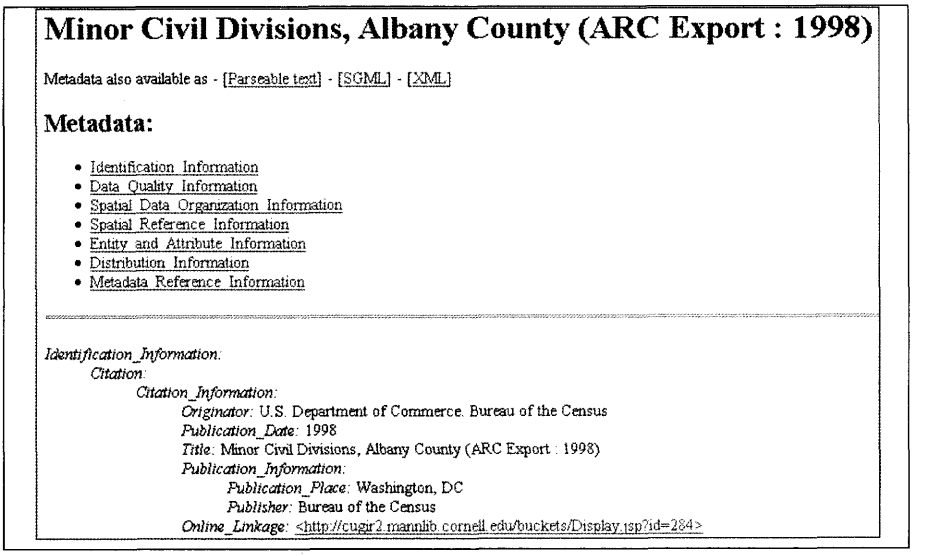
CUGIR Data

Currently CUGIR freely distributes online over 7,000 datasets produced by ten data partners, and their data come in seven unique proprietary and nonproprietary formats. [9] In many cases, one dataset is produced in multiple formats. For example, the dataset "Minor Civil Divisions, Albany County" is available in ArcExport as well as in shapefile format. Each format has unique characteristics that make it more or less desirable for certain uses and purposes. Unlike most digital library files that require little more than Internet connectivity and web browser software, geospatial data require technical expertise in the use of sophisticated and powerful GIS software applications. In addition, users must also understand cartographic and geographic concepts related to GIS.

CUGIR Metadata

In 1994 the Federal Geographic Data Committee established the Content Standard for Digital Geospatial Metadata for describing the content and function of geospatial data. There are 334 different elements in FGDC's CSDGM, 119 of which exist only to contain other elements.[10] These elements are organized within seven main sections and three supporting sections that describe different aspects of data that potential users might need to know: Identification Information, Data Quality Information, Spatial Data Organization Information, Spatial Reference Information, Entity and Attribute Information, Distribution Information, and Metadata Reference Information. For more extensive information about geospatial metadata, see Hart and Phillips's Metadata Primer.[11]

The Content Standard for Digital Geospatial Metadata is detailed, hierarchical, and complex. A high percentage of CUGIR geospatial metadata is provided by the data producer, and all of it is reviewed and enhanced by the metadata librarian to make it fully FGDC-compliant. Figure 9-1 is an example of a CUGIR record entitled "Minor Civil Divisions, Albany County." Note that the "Online_Linkage" element links users to the Dublin Core (DC) record where the data can be downloaded.



Minor Civil Divisions, Albany County (ARC Export : 1998)

Metadata also available as - [Parseable text] - [SGML] - [XML]

Metadata:

- Identification_Information
- Data_Quality_Information
- Spatial_Data_Organization_Information
- Spatial_Reference_Information
- Entity_and_Attribute_Information
- Distribution_Information
- Metadata_Reference_Information

Identification_Information:
Citation:
Citation_Information:
Originator: U.S. Department of Commerce. Bureau of the Census
Publication_Date: 1998
Title: Minor Civil Divisions, Albany County (ARC Export : 1998)
Publication_Information:
Publication_Place: Washington, DC
Publisher: Bureau of the Census
Online_Linkage: <http://cugir2.mannlib.cornell.edu/buckets/Display.jsp?id=284>

FIGURE 9-1 Geospatial/FGDC metadata record in CUGIR. From this record, one may download the dataset from the online linkage.

Of the 7,117 datasets in CUGIR, 7,111 are accompanied by FGDC-compliant metadata. CUGIR metadata are created and stored as ASCII text, HTML, SGML, and XML. Online users may view any metadata record in any syntax of their choice.

CUGIR METADATA MANAGEMENT

Today the term "metadata management" is increasingly being used by librarians, computer scientists, information scientists, and the e-commerce community.[12] Although libraries managed metadata long before it was known as metadata, the term "metadata management" has not been completely defined. Some practitioners indicate that it is an organizational process that can or cannot be automated, but the author takes the term one step further: "In a broad sense and in the case of CUGIR, metadata management implies the implementation of a metadata policy (i.e., principles that form the guiding framework within which metadata exists) and adherence to metadata standards."[13] Furthermore, metadata management is the process of acquiring and maintaining a controlled set of metadata, with or without automation, in order to describe, discover, preserve, retrieve, and access the data to which it refers.[14]

As problems arose in the development of CUGIR, it became clear that although the CUGIR team and its data partners had been creating metadata for years, there had never been a metadata policy that was explicitly articulated for them. This oversight was exposed when the CUGIR team began to approach preservation—since preservation policy should rest heavily on metadata policy. Although metadata policy and management are not panaceas for digital library woes, metadata management can ensure efficiency, interoperability, extensibility, and cost effectiveness through a clear and concise plan. The more complex, relational, and heterogeneous CUGIR metadata became, the more it became necessary to adopt a metadata policy as well as a preservation policy that would inform a metadata management system to deal with preservation, access, data and metadata versioning, and redundancy.

The CUGIR team identified one major area essential to CUGIR's success—access. It was clear to the team that Cornell University's core constituency of faculty, students, and staff were not sufficiently utilizing CUGIR's geospatial resources. In order to make geospatial information

resources more accessible to users who might not otherwise encounter them, CUGIR's FGDC records were converted to MARC and added to the library's online catalog and OCLC's FirstSearch. In addition, FGDC records were converted to Dublin Core (DC) and subsequently harvested by the Open Archives metadata harvester.[15]

Another identified problem was the prevalence of redundant metadata records that differed only in syntax, i.e., HTML or XML (Extensible Markup Language). The storage of metadata in HTML, XML, SGML, and ASCII text was difficult to manage when changes were necessary. Similarly, the repetition of metadata elements or fields in those metadata also demonstrated inefficient use of storage space. In order to address these problems, the CUGIR team set out to introduce a more accessible and efficient management system, based on the notion of one canonical metadata work.

Canonical CUGIR Metadata

In order to minimize the amount of data lost as a result of crosswalking among multiple schemas, the metadata schema-conversion process began with the core, or canonical, FGDC record that is assembled on-the-fly. The FGDC record is considered the "native" and most complete source of information in one of the most flexible exchange syntaxes, XML. With no existing tools to convert FGDC XML to MARC XML, this was quite a challenge. Elizabeth Mangan of the Library of Congress created an FGDC-to-MARC 21 crosswalk that was a useful beginning, but a new and customized FGDC XML-to-MARC XML crosswalk had to be created to suit our purposes.[16] The MARC XML is also derived from the canonical form and is produced on-the-fly.

What makes the use of the canonical record even more important is the upcoming introduction of International Organization for Standardization (ISO) geospatial metadata. The ISO metadata, when implemented, will harmonize the FGDC Metadata Standard (FGDC-STD-001-1998) with the ISO's Geographic Information/Geomatics Technical Committee (TC) 211 Metadata Standard 19115.[17] The standard will be expressed as a multilingual XML Schema designed to be extensible, multilayered, and modeled in Unified Modeling Language (UML).[18] In addition, it will be integrated with other ISO standards such as Dublin Core (ISO 15836:2003) and Codes for the Representation of Names of Languages (ISO 639-2).[19]

This harmonization process is a powerful step in the right direction because it not only addresses many known deficiencies in FGDC CSDGM, but also enables interoperability while providing additional support for the functions of metadata. Embracing XML-encoded FGDC is the CUGIR team's way of preparing for the upcoming changes. Given the metadata tools and practices we have in place, we expect a predictable and effortless transition from FGDC to ISO. Thus CUGIR will be poised to make an early transition, instead of waiting for proprietary metadata tools to emerge. The canonical record is stored in a database and is produced on-the-fly. This method allows for the introduction of some efficiencies; for example, each data partner has standard contact information (e.g., address, telephone number). Instead of repeating such information in each and every metadata record, it is stored once and rendered dynamically. Figure 9-2 illustrates the CUGIR metadata conversion process.

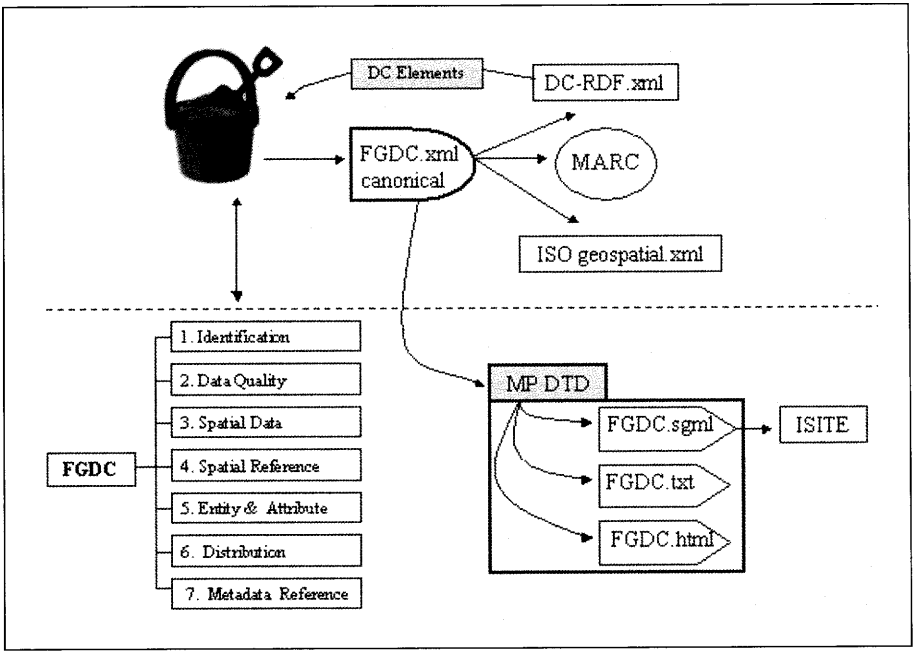


FIGURE 9-2 CUGIR metadata conversion process

As shown in the figure, the FGDC metadata is stored within a relational database and links to the bucket that is populated by Dublin Core. In addition, the activities above the line represent the new way of managing the metadata, and the activities below the line represent the old way of producing metadata records manually.

Resource Description Framework for Open Archives Initiative and the Semantic Web

The Open Archives Initiative (OAI) Metadata Harvesting Protocol was the only metadata-sharing tool, outside of CUGIR and the National Geospatial Data Clearinghouse, that was used to enhance access to CUGIR.[20] The minimum requirement for metadata in OAI is simple Dublin Core.[21] The CUGIR team chose to use the Resource Description Framework (RDF) for a number of reasons, the first being the convenient use of OCLC's Connexion to export OAI-ready DC in RDF with little effort.[22] As the metadata project progressed, we favored a less OCLC-centric approach to metadata creation. Moreover, we discovered that DC-compliant RDF records (in XML) could be easily created with XML stylesheets (XSL) coupled with Extensible Stylesheet Language Transformations (XSLT).[23] The use of RDF can be justified by its integral role in the Semantic Web.

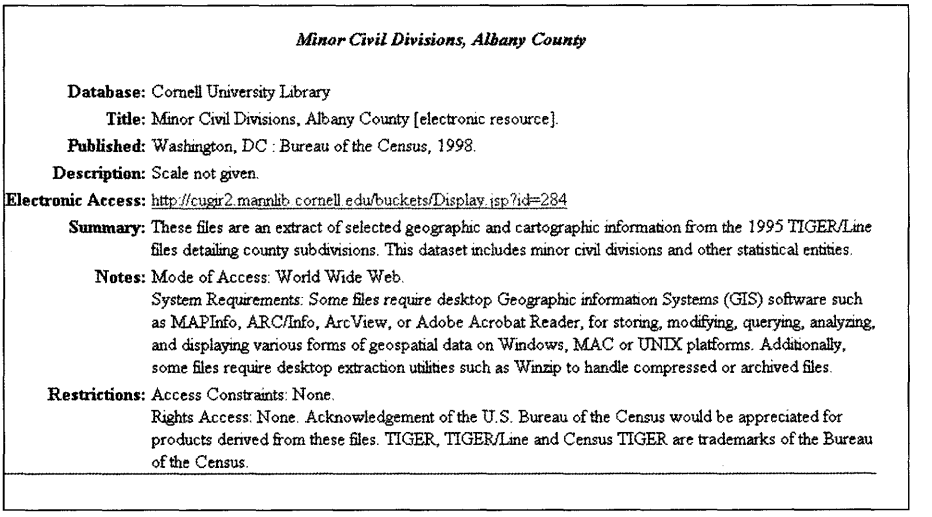
Metadata Management with MARC

The contribution of MARC 21 records to OCLC makes CUGIR data internationally accessible to WorldCat users. Additionally, other libraries on the OCLC network get the opportunity to utilize full-level MARC records. The integration of CUGIR data into the Cornell University OPAC made it possible for library users to discover geospatial resources as they typically discover journals, books, and online databases. In sum, the transformation from FGDC to MARC 21 enabled the CUGIR team to do the following:

1. Gain bibliographic control over CUGIR metadata records outside of CUGIR.
2. Enhance access to geospatial records via the OPAC.
3. Share MARC 21 records with libraries worldwide via WorldCat.

A CUGIR MARC 21 record is based on the XML-encoded FGDC records and transformed on-the-fly using XSLT. See figure 9-3 for an example of a MARC 21 record in the Cornell University Library's OPAC based on the FGDC record shown in figure 9-1.

While the team was already creating multiple metadata schemas, it seemed only natural to include some of the latest developments in metadata, such as the Metadata Object Description Schema (MODS).[24] The addition of MODS into the metadata framework forced the team to create an FGDC-to-MODS crosswalk, stylesheet, and transformation, since none existed.[25] The MODS schema is a flexible XML-based descriptive standard which can be combined with other XML-based standards, including the Metadata Encoding and Transmission Standard (METS).[26] METS, a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, fills in essential components needed to manage a digital library. Since any descriptive metadata that is part of CUGIR can be part of METS objects, we anticipate that the next step will be to investigate how well METS can handle geospatial information.



Minor Civil Divisions, Albany County

Database: Cornell University Library

Title: Minor Civil Divisions, Albany County [electronic resource].

Published: Washington, DC : Bureau of the Census, 1998.

Description: Scale not given.

Electronic Access: http://cugir2.mannlib.cornell.edu/buckets/Display.jsp?id=284

Summary: These files are an extract of selected geographic and cartographic information from the 1995 TIGER/Line files detailing county subdivisions. This dataset includes minor civil divisions and other statistical entities.

Notes: Mode of Access: World Wide Web.
System Requirements: Some files require desktop Geographic information Systems (GIS) software such as MAPInfo, ARC/Info, ArcView, or Adobe Acrobat Reader, for storing, modifying, querying, analyzing, and displaying various forms of geospatial data on Windows, MAC or UNIX platforms. Additionally, some files require desktop extraction utilities such as Winzip to handle compressed or archived files.

Restrictions: Access Constraints: None.
Rights Access: None. Acknowledgement of the U.S. Bureau of the Census would be appreciated for products derived from these files. TIGER, TIGER/Line and Census TIGER are trademarks of the Bureau of the Census.

FIGURE 9-3 MARC record in the Cornell University Library online catalog. Notice how the Electronic Access (MARC 856) field is identical to the link, no. 284, found in figure 9-1.

Metadata Editing and Automatic Metadata Creation and Synchronization

CUGIR currently uses a suite of software produced by the Environmental Systems Research Institute (ESRI), commonly used in geospatial information analysis, to manage and store CUGIR data and metadata. These include the software components ArcGIS, an Internet Mapping Service (ArcIMS), and a Spatial Data Engine (ArcSDE). ArcGIS contains a data management tool known as ArcCatalog, which is a data exploration and management application used to preview metadata as well as a dataset's geographic and tabular data. It automatically creates metadata for datasets stored in the geodatabase if none exists. Some of the automatically generated metadata describe the data's current properties, i.e., coordinate system, entity, and attribute information. Every time the metadata librarian views the metadata, ArcCatalog automatically updates or synchronizes dataset properties with its most current values. The synchronization ensures that the metadata is perpetually up-to-date according to the changes in the dataset. Automatic synchronization is invaluable, but it brings forth a host of problems associated with archiving and bibliographic control. Making distinctions between and among metadata versions, editions, and updates is crucial for any type of digital library with archiving responsibilities such as CUGIR. The inability of the synchronizer to differentiate a version of a metadata record from an edition or update brought forth a new set of challenges.

FUNCTIONAL REQUIREMENTS FOR BIBLIOGRAPHIC RECORDS OVERVIEW

When the CUGIR team needed to determine the key issues in distinguishing and classifying CUGIR metadata, it was clear that the FRBR entity hierarchy could provide some guidance. CUGIR, like most digital libraries, organizes data linearly. There is a one-to-one relationship between CUGIR datasets and metadata. The metadatabase system in ArcCatalog displays bibliographic information in hierarchical ways, yet the a priori relationships are not fully captured. Fortunately, CUGIR's Smart Object Dumb Archive (SODA) architecture alleviates the problem by displaying alternate expressions of datasets, but SODA cannot fully capture the hierarchical relationship inherent to the data. The intricate

details of the SODA model have been well documented by its creator, Michael Nelson.[27] The similarities and differences among expressions, manifestations, and items pose unique challenges for the archiving, preservation, and organization of CUGIR data.

In some cases, changes to the intellectual content of the dataset (e.g., datum) are reflected in its respective metadata. Similarly, a change in the way a particular dataset is packaged (e.g., compression) can also be handled under synchronization. On the other hand, there are often changes to the data that are not necessarily recognized by the synchronizer. For example, a change in a keyword would not be apparent to the metadata synchronizer, but represents nonetheless a key access point change in the metadata. The CUGIR team works frequently with data partners that are more familiar with the world of GIS than with theory and research regarding the intellectual organization of information. Geospatial information practitioners do not make distinctions between intellectual content and physical packaging, but in the world of libraries such issues are viewed as critical. These relationships, nuances, and embodiments of CUGIR metadata records should be examined under the FRBR lens in order to secure clarity over what should be and should not be synchronized.

FRBR and CUGIR Metadata

The FRBR model can assist in determining what should be the appropriate unit of storage for the organization, discovery, preservation, and description of CUGIR data. Any substantial changes to the canonical FGDC record means that the derivative records (DC-RDF, MARC 21, MODS) must be changed as well. The design of the CUGIR metadata model is in concert with Jenkins et al.'s assertion: "Automatic metadata generation would appear to be an essential pre-requisite for widespread deployment of RDF based applications."[28] The application of the FRBR model to CUGIR records is shown in figure 9-4.

The CUGIR team is still negotiating methods by which the synchronizer can be programmed to form an FRBR-like hierarchy when metadata needs to be changed. Since the synchronizer does not understand the difference between an intellectual and physical change, the metadata records were parsed in such a way as to require a command that dictates: <when field 1.1.2 (thesaurus field) changes in FGDC record, do not synchronize metadata because it has intellectually changed>. Although the entire

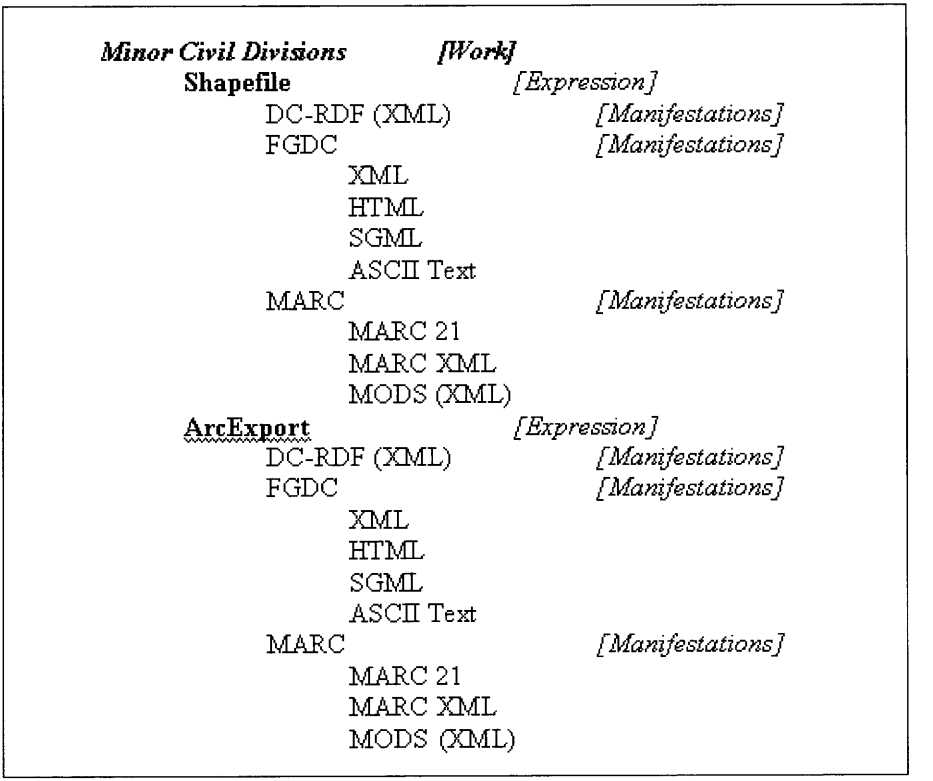


FIGURE 9-4 CUGIR metadata conceptualized in the FRBR work entity framework

analysis of CUGIR data is incomplete, it is clear that CUGIR data does not fit neatly into the FRBR model.

LESSONS LEARNED

During the course of any metadata-intensive project, the tools (software), knowledge, and the metadata schemas will change. In hindsight, there is little that the CUGIR team could have done to improve the metadata management model. This is because changes to the software, the team's knowledge set, and the metadata standards happened unpredictably throughout the implementation of the metadata management system.

Metadata. When the project began, CUGIR utilized the existing metadata standard, FGDC. Currently, ISO metadata, in an XML Schema, has been approved and destined to replace FGDC. This transition from FGDC to ISO was one of the biggest catalysts that forced the team to expand their use of metadata standards.

Software. The software and tools that were developed for the project changed as the standards and understanding changed. When the CUGIR metadata management project was conceived, it was designed to deal with metadata in SGML (Standard Generalized Markup Language), not XML. Because CUGIR was using the Isite software, which required SGML, the team was working with the assumption that ISITE and SGML, respectively, would be used for indexing CUGIR metadata for three more years.[29] It became clear that SGML was too cumbersome, so the team was forced to re-create the tools with XML in mind. In addition, the CUGIR data was migrated to the proprietary software package produced by the GIS leader ESRI, eliminating the last remaining need for SGML.

Knowledge. Probably the most important and underestimated factor that had an impact on the progress of the project was the knowledge base of the team. As the programmer and the librarians involved became more knowledgeable about the utility of RDF, their ideas began to shift. The placement of RDF within the model happened as the team became exposed to more information about RDF, the Semantic Web, and ontologies.

OUTCOMES OF CUGIR METADATA FRAMEWORK

The CUGIR metadata framework proved successful in reaching its primary goals: increasing access and implementing an efficient metadata management system. But clearly the test of the system's effectiveness is in the question of whether more users discovered CUGIR as a result of the metadata framework.

When the framework was implemented, referrer data, which indicated the web page that a user visited in order to access the bucket, was captured and stored in a database. The IP addresses of the hosts were also collected. To preserve the privacy of users, the IP addresses were encrypted and the subnets dropped from the statistics database. As a result, the domain name rather than the unique address of the computer has been stored. These data identify whether users encountered a bucket from OAI, the

Cornell OPAC, or OCLC's FirstSearch as their entry into CUGIR. Since the metadata framework has been in place over 12,000 buckets have been accessed from a variety of locations. The results indicate that less than 5 percent of our users discover CUGIR metadata via the Cornell OPAC. Less than one percent of our users discover CUGIR metadata via FirstSearch. Almost 95 percent of our users discover CUGIR metadata from CUGIR's home page.

If only 5 percent of our users discovered CUGIR as a result of this metadata framework, was it worthwhile? Although the statistics do not indicate "success" in regard to access, the work and process of formulating the metadata-sharing framework forced us to document all metadata processes, streamline workflows, and create more metadata with less effort. In terms of data management, the metadata framework reduced the number of metadata files that had to be managed and stored. CUGIR no longer stores each metadata schema in multiple formats. In the past, we stored nine metadata files per dataset; now we only store one.

CONCLUSION

We are confident that our work to make CUGIR more accessible will pay off in the long run. Furthermore, the proliferation of web mapping services will expose GIS data to even more users. Increasingly diverse and sophisticated interactive mapping websites, allowing instant creation of customized maps, exemplify the most dynamic aspects of GIS usage. Many repositories are beginning to offer interactive mapping websites where one can create maps based on large census, Environmental Protection Agency, or U.S. Geological Survey databases of information.

Finally, the value of the CUGIR metadata framework is promising when one examines the growing importance of standards in the GIS community. Consortia such as the Open GIS Consortium are aimed at growing interoperability for technologies involving spatial information and location, so that benefits from geographic information and services can be made available across any network, application, or platform.[30] With this in mind, analysis of data on the use of the CUGIR metadata management system yields some interesting insights:

1. In spite of the vast efforts to make CUGIR data accessible across metadata schemas and information systems, users who know

about CUGIR overwhelmingly prefer to acquire data from the FGDC metadata records on the CUGIR home page. This may always be the case no matter how much metadata sharing persists.
2. The OPAC provides discovery but minimal means for access for users who might not otherwise discover geospatial data.
3. The addition of MARC 21 records in OCLC has not significantly increased access to CUGIR, but other libraries in the OCLC network have access to full-level MARC records and may find them useful.
4. The application of the FRBR model helped the team make clearer distinctions among metadata surrogates, but it did not necessarily solve the problems that GIS software presents to digital libraries.

The fundamental value of the library is the organization of information as the foundation through which information resources can be utilized. Centuries of library research support this claim. The same principles are not routinely being applied to digital libraries. The CUGIR team embraces metadata as the first-order prerequisite to establishing a complete geospatial repository. Furthermore, it should be clear that library standards and theory as well as GIS standards and software must be applied in concert, in order to produce open, interoperable, efficient, and robust digital libraries.

NOTES

1. Cornell University Geospatial Information Repository (home page), http://cugir.mannlib.cornell.edu (accessed 16 December 2003).
2. Geospatial data are typically born digital and by definition are digital representations of real-world features that describe objects and relations among them. Additionally, GIS data include spatial reference information, contain both geometric and thematic data, come in many different formats, support a wide range of applications, and offer more flexibility than hard-copy maps. When it became clear that GIS data was being lost (e.g., the 1968 New York Land Use Study), the CUGIR team looked to long-standing library standards and practices in preservation to prevent more data loss. The CUGIR team was unaware of major preservation initiatives, in part because GIS practitioners are more focused on building resources than on saving older versions of data for posterity. For many GIS practitioners, GIS data are ephemeral, but for libraries, preservation is paramount. Through our discussions about preserving CUGIR data, we began to focus on what makes GIS data unique and problematic. Some of the information gathered about GIS data was presented at Cornell University Library's conference on "Digital Preservation Management: Implementing Short Term Strategies for Long Term Problems," Ithaca, N.Y., August 2003.

3. By "editioning," I refer to metadata records that are updated and become first, second, or third editions of the same work. This concept must be distinguished from "versioning," which typically pertains to the information that accompanies the format of the data that the metadata is describing.
4. IFLA Study Group on the Functional Requirements for Bibliographic Records, *Functional Requirements for Bibliographic Records: Final Report* (Munich: K. G. Saur, 1998).
5. Fixity is used to ensure that the particular content information object has not been altered in an undocumented manner, according to "Preservation Metadata and the OAIS Information Model, A Metadata Framework to Support the Preservation of Digital Objects: A Report by The OCLC/RLG Working Group on Preservation Metadata," June 2002, http://www.oclc.org/research/projects/pmwg/pm_framework.pdf (accessed 23 November 2003). To quote John Kunze, researcher at the University of California at San Francisco's Library Center for Knowledge Management: "Permanence of electronic information, namely, the extent to which structured digital data remains predictably available through known channels, is a central concern for most organizations whose mission includes an archival function" (http://www.nii.ac.jp/dc2001/proceedings/product/paper-27.pdf; accessed 15 December 2003). Michael Nelson and B. Danette Allen also talk about object persistence in their article, "Object Persistence and Availability in Digital Libraries," *D-Lib Magazine,* January 2002, http://www.dlib.org/dlib/january02/nelson/01nelson.html (accessed 15 December 2003).
6. Philip Herold, Thomas D. Gale, and Thomas Turner, "Optimizing Web Access to Geospatial Data: The Cornell University Geospatial Information Repository," *Issues in Science and Technology Librarianship* (winter 1999), http://www.library.ucsb.edu/istl/99-winter/article2.html (accessed 24 January 2003).
7. Herold, Gale, and Turner, "Optimizing Web Access."
8. CUGIR, "CUGIR Statistics Database," http://cugir.mannlib.cornell.edu/about_cug/ (accessed 7 March 2004).
9. As of 2003, CUGIR file formats included ArcExport, shapefile, CAD, geoTIFF, PDF, ArcInfo Grid, and DEM. More information about them can be found at "CUGIR Help Files," http://cugir.mannlib.cornell.edu/help/help.html (accessed 5 April 2003).
10. Peter Schweitzer, "Frequently Asked Questions on FGDC Metadata," http://geology.usgs.gov/tools/metadata/tools/doc/faq.html (accessed 2 February 2003).
11. D. Hart and Hugh Phillips, "Metadata Primer—'How To' Guide on Metadata Implementation," 2001, http://www.lic.wisc.edu/metadata/ metaprim.htm (accessed 10 August 2001).
12. OCLC: Online Computing Library Center, "Metadata Management and Knowledge Organization," 2003, http://www.oclc.org/research/projects/metadata_management.htm. (accessed 19 December 2003). In the private sector, a host of companies are devoted to this endeavor, including Metadata Management Corp. and Agilense, Inc. See also Diane I. Hillmann, "Metadata Management," 2002, http://metamanagement.comm.nsdl.org/cgi-bin/wiki.pl (accessed 23 November 2003).

13. Elaine L. Westbrooks, "Distributing and Synchronizing Heterogeneous Metadata for the Management of Geospatial Information in DC-2003," in *Proceedings of the International DCMI Metadata Conference and Workshop,* Seattle, Wash., 28 September–2 October 2003, http://www.siderean.com/dc2003/204_Paper78.pdf (accessed 7 March 2004).
14. Intra-Governmental Group on Geographic Information Working Group, "Principles of Good Metadata Management," 2002, in *Intra-Governmental Group on Geographic Information, Working Group on Metadata Implementation Guide,* http://www.iggi.gov.uk/achievements_deliverables/pdf/Guide.pdf (accessed 17 July 2003).
15. Carl Lagoze et al., "The Open Archives Initiative Protocol for Metadata Harvesting," 2001, http://www.openarchives.org/OAI/openarchivesprotocol.html (accessed 1 March 2003).
16. Elizabeth Mangan, "Crosswalk: FGDC Content Standards for Digital Geospatial Metadata to USMARC," 1997, http://alexandria.sdc.ucsb.edu/public-documents/metadata/fgdc2marc.html (accessed 12 December 2000); Elaine L. Westbrooks, "FGDC Content Standards for Geospatial Metadata to MARC and MODS Crosswalk," 2003, http://metadata-wg.mannlib.cornell.edu/elaine/fgdc/fgdc2mod4.html (accessed 7 March 2004).
17. Federal Geographic Data Committee, "FGDC/ISO Metadata Standard Harmonization," 2003, http://www.fgdc.gov/metadata/whatsnew/fgdciso.html (accessed 4 November 2003).
18. Federal Geographic Data Committee, "FGDC/ISO Metadata Standard Harmonization."
19. The Dublin Core Metadata Element Set ISO 15836:2003(E) was approved to be an ISO standard on 4 August 2003. See http://www.niso.org/international/SC4/n515.pdf (accessed 15 December 2003); Library of Congress, Network Development and MARC Standards Office, "Codes for the Representation of Names of Languages—Part 2," 2003, http://www.loc.gov/standards/iso639-2/ (accessed 15 December 2003).
20. According to the Federal Geographic Data Committee, the National Geospatial Data Clearinghouse "is a collection of over 250 spatial data servers that have digital geographic data primarily for use in GIS, image processing systems, and other modelling software. These data collections can be searched through a single interface based on . . . metadata." See http://130.11.52.184 (accessed 16 December 2003).
21. Carl Lagoze et al., "Open Archives Initiative Frequently Asked Questions," 2002, http://www.openarchives.org/documents/FAQ.html (accessed 1 March 2003).
22. Connexion is OCLC's online cataloging service that is used to create and edit bibliographic and authority records, as well as harvest metadata from online resources. For more information, see OCLC: Online Computing Library Center, "Connexion—Cataloging and Metadata," 2003, http://www.oclc.org/connexion/ (accessed 15 December 2003). For more information about Dublin Core expressed within the Resource Description Framework (DC-RDF), see Dave Beckett, "Expressing Simple Dublin Core in RDF/XML," http://dublincore.org/documents/2002/04/22/dcmes-xml/index.shtml (accessed 15 December 2003). See also Stefan Kokkelink, "Expressing Qualified Dublin Core in

RDF/XML," http://dublincore.org/documents/2001/08/29/dcq-rdf-xml/index.shtml (accessed 12 December 2003).

23. Extensible Stylesheet Language (XSL) defines how data are presented, while Extensible Stylesheet Language Transformations (XSLT) is designed for use as part of XSL.
24. Library of Congress, Network Development and MARC Standards Office, "MODS: The Metadata Object Description Schema" (home page), http://www.loc.gov/standards/mods/ (accessed 12 February 2003). Rebecca Guenther, senior networking and standards specialist at the Library of Congress, adds: "MODS should complement other metadata formats and should provide an alternative between a very simple metadata format with a minimum of fields and no or little substructure [i.e., Dublin Core] and a very detailed format with many data elements having various structural complexities such as MARC 21."
25. Elaine L. Westbrooks, "FGDC to MODS Crosswalk," 2003, http://metadata-wg.mannlib.cornell.edu/elaine/fgdc/ (accessed 30 April 2003).
26. Library of Congress, "METS: Metadata Encoding and Transmission Standard," http://www.loc.gov/standards/mets/ (accessed 6 May 2003).
27. Michael L. Nelson, "Buckets: Smart Objects for Digital Libraries," 2000, unpublished Ph.D., Old Dominion University, Norfolk, Va.; Michael L. Nelson, "Smart Objects and Open Archives," *D-Lib Magazine,* February 2001, http://www.dlib.org/dlib/february01/nelson/02nelson.html (accessed 15 January 2002); Michael L. Nelson, "Smart Objects, Dumb Archives: A User-Centric, Layered Digital Library Framework," *D-Lib Magazine,* March 1999, http://www.dlib.org/dlib/march99/maly/03maly.html (accessed 19 December 2003).
28. Charlotte Jenkins et al., "Automatic RDF Metadata Generation for Resource Discovery," *Computer Networks* 31 (1999): 1305–20.
29. Federal Geographic Data Committee, "Federal Geographic Data Committee FAQs: What Is the Isite Software and What Do Each of the Components Do?" 2003, http://clearinghouse4.fgdc.gov/fgdcfaq/showquestion.asp?faq=3&fldAuto=13 (accessed 16 December 2003).
30. Open GIS Consortium, "About the Open GIS Consortium," 2003, http://www.opengis.org/ogcAbout.htm (accessed 2 May 2003).